

DOCUMENT RESUME

ED 038 385

TE 001 042

AUTHOR Carruthers, Robert B.
TITLE Building Better English Tests: A Guide for Teachers of English in the Secondary School.
INSTITUTION National Council of Teachers of English, Champaign, Ill.
PUB DATE 63
NOTE 31p.
AVAILABLE FROM National Council of Teachers of English, 508 South Sixth Street, Champaign, Ill. 61820 (Stock No. 00607, \$1.00, prepaid)

EDRS PRICE EDRS Price MF-\$0.25 HC-\$1.65
DESCRIPTORS *Achievement Tests, *English Instruction, Essay Tests, Multiple Choice Tests, Objective Tests, Student Evaluation, *Student Testing, *Test Construction, *Testing, Testing Problems, Test Interpretation, Test Reliability, Test Selection

ABSTRACT

The primary purpose of this leaflet is to assist the teacher of English in building good achievement tests, with special attention to planning and construction. Subjects covered are general considerations for planning a test, making a blueprint for the test, basic characteristics of effective tests and test questions, selecting the proper test questions, building effective short-answer and essay questions, and reviewing the test. (LH)

ED038385

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE
PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION
POSITION OR POLICY.

BUILDING BETTER ENGLISH TESTS

**A Guide for Teachers of English
in the Secondary School**

Robert B. Carruthers

TE 001 042

NATIONAL COUNCIL OF TEACHERS OF ENGLISH

508 South Sixth Street

Champaign, Illinois

ED038382

CONSULTING READERS OF THE MANUSCRIPT

Harold D. Carter, University of California, Berkeley

William H. Evans, University of Illinois, Urbana

David H. Russell, University of California, Berkeley

Copyright 1963

National Council of Teachers of English

"PERMISSION TO REPRODUCE THIS
COPYRIGHTED MATERIAL HAS BEEN GRANTED

BY NCTE and author

TO ERIC AND ORGANIZATIONS OPERATING
UNDER AGREEMENTS WITH THE U.S. OFFICE OF
EDUCATION. FURTHER REPRODUCTION OUTSIDE
THE ERIC SYSTEM REQUIRES PERMISSION OF
THE COPYRIGHT OWNER."

NCTE COMMITTEE ON PUBLICATIONS

James R. Squire, NCTE Executive Secretary, Chairman

Jarvis E. Bush, Wisconsin State College, Oshkosh

Autrey Nell Wiley, Texas Woman's University, Denton

Miriam E. Wilt, Temple University, Philadelphia

Enid M. Olson, NCTE Director of Publications

540 100 54

1963 100 54

TABLE OF CONTENTS

I. AN OVERVIEW	5
II. PLANNING THE TEST: GENERAL CONSIDERATIONS	6
III. PLANNING THE TEST: MAKING THE BLUEPRINT	7
IV. BASIC CHARACTERISTICS OF EFFECTIVE TESTS AND TEST QUESTIONS	9
V. SELECTING THE PROPER TEST QUESTIONS: GENERAL CONSIDERATIONS	11
VI. BUILDING EFFECTIVE SHORT-ANSWER ITEMS	13
VII. BUILDING EFFECTIVE ESSAY QUESTIONS	25
VIII. REVIEWING THE TEST	30
IX. BIBLIOGRAPHY	32

I. AN OVERVIEW

The increasing importance of tests in the lives of high school students is evident in the emphasis given to tests which are prepared, standardized, and administered on a national basis. In some instances, tests are used as a basis for admission to a college or university, or as a predictor of college success; in others, tests indirectly influence the high school curriculum, teaching practices, and testing program. Thus, tests prepared by agencies outside of the high school have an impact upon teaching. However, it is the classroom test constructed by the teacher which is the type most commonly used in schools today. In English, as in other subjects, the classroom test remains the backbone of the school's testing program in the subject.

Teachers of English, thoroughly competent in scholarship and in teaching skill, are taking their cue from makers of standardized tests. They are paying increasing attention to the tests which they build to measure student achievement. And they are writing more effective tests. English teachers wish to measure, among other objectives, the acquisition of knowledge, reading ability, and the more complex abilities of mature appreciation, judgment, critical thinking and reading, response to subtleties of expression, response to symbolism in literature, and interpretation of data. These latter emphases bespeak an increasing sensitivity on the part of many teachers to the more intangible outcomes of a program in English. How to reflect these emphases in test items is a problem which many teachers wish to explore.

All teachers of English, however, are concerned with building effective tests. Two of the areas in which teachers have expressed a desire to strengthen their work are (1) the planning of a test and (2) the construction of valid, reliable test questions. The principal purpose of this leaflet, then, is to assist the teacher of English in building good achievement tests, with especial attention to planning and construction. The leaflet presents an analysis of the procedure for building a test and contains suggestions, techniques, and cautions for building test items, together with illustrative items and suggestions for reviewing a test. The author hopes that the leaflet will help the teacher to become a more skilled practitioner of the art—as well as the science—of testing.

For purposes of convenience, the author, in the early sections (II, III, IV, and V) of the leaflet, describes the processes involved in making a test for a unit—"Understanding Human Nature"—for an English 10 (sophomore) class. He follows this procedure because many teachers teach units of work and because the principles involved in the construction of a good unit test are inherent in those for building other types of tests. The other sections of the leaflet deal with principles for constructing test items for the sample unit test. To make the leaflet as helpful as possible, the author also has included items to illustrate other principles for building items, and a concluding section (VIII) on reviewing and editing a test.

Although the content of this leaflet is limited to paper and pencil examinations for classroom use, the teacher of English will find that other means of evaluation, such as observation, interviews, anecdotal records, and oral tests, also are very useful.

II. PLANNING THE TEST: GENERAL CONSIDERATIONS

Classroom tests in English, as in other subjects, have certain outstanding uses for the teacher. They can serve to stimulate the student to achieve, to motivate the student, to evaluate student achievement, to diagnose student abilities, to suggest needed areas for additional work, and to some extent to suggest the effectiveness of the course. They can help to determine those final marks. Also, a teacher can readily adjust a classroom test to the needs of a particular class, whereas a standardized test cannot be changed. Furthermore, if well constructed, a teacher-made test can reflect better than a standardized test the particular emphasis that has been given to the various areas of English in a course.

Like any test, a unit test in English is a *sample* of the content of the unit. As the teacher begins to build the test, he must take several important steps. In effect, he must *plan the test*.

The teacher must first (1) review his objectives for the unit. The test should measure *all* of the important objectives of the unit. For example, for our sample unit for English 10, "Understanding Human Nature," the teacher may have for one of his objectives "To enable the students to become more aware of characters' motives." The teacher should review this objective and the steps he took to attain it. He should then (2) decide whether this objective can be effectively tested in a paper and pencil test, and (3) decide what kind of written evidence will reflect the student's having attained the objective. He decides that it can be tested and concludes that such specific evidence as this will suffice: "Awareness of the effect of a character's actions upon others"; "awareness of the implications of a character's conversations with others"; and "understanding of the inner conflict of a character." Thus, he has sharpened the *general* teaching objective ("To enable the students to become more aware of characters' motives") to the specific *testing* ones ("Awareness of the effect of a character's actions upon others . . .," etc.). He must then in the same manner point up and sharpen each of the other general objectives for the unit. This is a challenging task but one that pays dividends later.

The teacher needs to make each testing objective as *specific* as possible, in the manner of the above example. (He is now beginning to make the blueprint below.) For example, in our sample unit, "The ability to punctuate correctly" is, for most testing purposes, a rather broad testing objective. "The ability to use the comma" is a somewhat more specific one; however, it too is rather broad. [To use the comma when? in what situations? to reveal ability to use what rules?] Let us try again. "The ability to use the comma before 'and,' 'but,' 'for,' 'nor,' when used as coordinating conjunctions joining two independent clauses" is a satisfactory specific testing objective. Again, such an objective as "The ability to use the library card catalog" can be reduced to "The ability to use a subject heading card in the library card catalog," if this is the specific skill to be measured. In each of the above cases, the latter objective is much more specific than the former. If the teacher constructs questions which will measure the specific objective, his test will tend to have more validity, a property of tests which we shall discuss later.

It is also quite possible that the teacher, in reviewing the objectives of the unit and in planning the test, will gain a new perspective on the content and emphases of the unit and on what he wishes to test. Deciding what can be tested and sharpening it to specific testing objectives will help him in evaluating the unit. He may even say to himself, "I certainly stressed—(objective), but I really ignored—(objective)." Such an "agonizing reappraisal" can be salutary.

III. PLANNING THE TEST: MAKING THE BLUEPRINT

Having selected and defined the objectives to be tested, the teacher should next complete a *blueprint* for the test. As the name implies, a blueprint is a plan. The blueprint below is a plan for a test on our unit on "Understanding Human Nature." It includes all of the objectives as sharpened and, in addition, reflects the emphasis which the teacher assigns to each of the objectives to be measured by the test. He gives each of these objectives a *weight*.

How does he derive each weight? For each objective, he assigns a weight commensurate with the emphasis he accorded it during the course of the unit. In the interests of fairness and validity, he will make sure, for example, that if in our sample unit he spent about 50 percent of the time—and emphasis—of the unit on the study of literature and its outcomes, he will allocate on the test a like weight (50 percent) for items in the area of literature and its outcomes. Again in this example, to stress literature work for 50 percent or more of the work of the unit and then to attach a weight of only 20 percent in the test is unfair and unrealistic. So are similar inequities in the converse of this example.

Note, too, that each weight is subdivided into *credits*, which also should reflect to some extent the previous emphasis of the unit. Here again the teacher should be careful to make a fair allocation, in the same manner as that of weight.

Upon completion of the blueprint, the teacher should review his plan for the test. Here are some important questions to be answered:

Are *all* important objectives of the unit being measured by the test?

For example, are objectives in appreciating, analyzing, and judging included in the unit?

If so, does the blueprint make provision for them?

Are the objectives being measured to the proper extent?

Is each objective with its subdivisions defined as accurately and specifically as possible—in words which exactly convey the idea?

For example, regarding word study, which specific ability of the student is being measured: To *know* the words? To *use* the words? To know the derivation of words? To recognize synonyms for a word? In other words, exactly what is being tested?

The teacher should consider each of these points carefully before he proceeds further. He will thus largely eliminate any subsequent problems, after the test is administered, about proper sampling of the content of the unit and about the specific ability, skill, or point which he wishes to test. As we shall

BLUEPRINT FOR A UNIT TEST
"Understanding Human Nature"

<i>Objective</i>	<i>Weight</i>	<i>Credits</i>
Understanding of characters' motives	50	
Awareness of the effect of a character's actions upon others		(10)
Awareness of the implications of a character's conversations with others		(20)
Understanding of an inner conflict of a character		(20)
Knowledge of, and use of, the card catalog to locate books of biography	10	
Ability to use a subject heading card		(6)
Knowledge of the Dewey Decimal System of Classification number for books of biography		(3)
Knowledge of the proper procedure for taking out a book from the library		(1)
Knowledge of, and use of, new words encountered in the unit	10	
Knowledge of the meanings of the new words		(5)
Ability to use, in writing, the new words		(5)
Ability to use the comma before "and," "but," "for," "or," "nor," when used as coordinating conjunctions joining two independent clauses	10	(10)
Ability to use, in writing, adjectives and adverbs to describe people encountered in the unit	10	(10)
Ability to spell new words encountered in the unit	10	(10)
Total	<u>100</u>	<u>(100)</u>

see, a careful blueprint will help him immeasurably in deciding what kind of test question to use.

It should be obvious, also, that the teacher can change a blueprint to meet the demands of different classes and groups. The demands may vary according to objectives, weights and credits, and testing techniques; that is as it should be. Also, in some situations, the teacher will wish to prepare an entirely new blueprint in order to achieve his purposes in testing.

IV. BASIC CHARACTERISTICS OF EFFECTIVE TESTS AND TEST QUESTIONS

Once the teacher has made his blueprint for the test, he is ready to consider the manner in which he can best measure the attainment of the objectives in the blueprint. Certain factors of tests and test questions greatly affect this goal. Let us review briefly some of the basic characteristics of effective tests and test questions (or *items*, as they are sometimes called).

A. Validity

The validity of a test or test item is the extent to which it actually measures what the teacher intends it to measure. It should be noted here that a test or test item is valid only for a *particular purpose*.

A first aspect of validity is curricular validity (in our case, validity for the unit). A test is said to have curricular validity if it reflects the goals and activities of the instruction, presents an adequate sample of content, and tests all of the important objectives of the instruction. In our sample unit, our blueprint reflects curricular validity.

A second aspect of validity is logical validity. For example, to test the student's ability to write correct English it is most probable—and logical—that some form of exercise in writing will yield the most conclusive evidence. Again, in our sample blueprint in Section III, to "test the student's ability to spell words encountered in the unit," an item for correction, in which the student *corrects* a misspelled word, is probably more valid for the purpose than a four-choice item in which the student is required merely to *select* a misspelled word but not to spell it correctly. (He may know which one of the four choices is misspelled, but he may not be able to *spell* it correctly.) Or, to test the student's "ability to use a comma before (certain) coordinating conjunctions which join two independent clauses," a short exercise which reads, "Punctuate correctly each of the following sentences" is probably more valid for the purpose than an item which reads, "Make a list of the uses of a comma." Again, in testing reading comprehension, an item may contain a word which causes a vocabulary problem for the student. This condition reduces the validity of the item because it is not reading ability, but to some extent vocabulary, which is being tested. In other words, each item should be built for a specific purpose and should test student achievement in terms of that purpose only. If the item does measure other aspects, they should be ignored in the scoring or grading of the answer.

A third characteristic of validity is the discriminatory power of an item. This is an imposing term: all it means is the ability of an item to separate the superior students from the poor students. If an item possesses good discrimination, superior students will achieve high scores or grades on it, and poor students will achieve low scores or grades. However, if there is no significant difference between the achievement of the two groups on the item, it probably has poor discrimination. Thus, a short-answer item will be more valid if it is carefully constructed along lines suggested in Section VI of this leaflet. Essay questions are more valid if they are carefully constructed and graded and if analysis of qual-

ities actually revealed by the answers shows definitely superior work by superior students and poor work by poor students.

B. Reliability

The reliability of a test or test item is its consistency, the extent to which the test gives the same results when repeated with the same or similar groups. The grade a student receives on a test should not be the result of his "misinterpretation" of a poorly worded question (with consequent loss of credit) or such practices as inconsistent scoring or grading of his answer by the teacher on different occasions. Certain types of essay questions are especially subject to this unreliability of grading. (See page 27.)

The reliability of a test or test item can be improved by making its requirements as unambiguous as possible. In a short-answer item, for example, the student should be completely clear as to what the teacher wants to know and as to the manner in which he is to indicate his answer. There should be only one correct answer for an item (or a minimum of correct answers, if more than one answer is eventually allowed). In an essay question, the student should have no difficulty in understanding the limits of the question and the directions for answering it—even if he cannot do so. The teacher can also improve the reliability of an essay question by setting up a key before he grades answers. This key should contain a description of the content, qualities, etc., desired in an acceptable answer, together with weights, as in the sample blueprint in Section III.

Each of the two qualities above—validity and reliability—particularly affects the various types of test items which are described in Sections V, VI, and VII. Specific examples of this effect are included in these sections, together with suggestions for improving the items especially insofar as these qualities are concerned.

C. Ease of Administration and Scoring

A test is said to be easily administered when certain criteria are met. The directions for answering test items are as clear and direct as possible. The student knows how and where to place his answer to a short-answer item (usually a space in the right margin is most feasible for scoring), and he understands the manner in which he is to answer an essay question. In the latter, he is provided additional scrap paper if necessary. The amount of total time for the test, as well as the allocation of credits, is adequate for the number and difficulty of the items. (Although it is difficult to estimate the total time, the teacher can reach a valid judgment if he bases it on his past experience or on the experience of others responsible for the construction of similar questions on a nationwide basis.) Where necessary, word and space limits are sharply defined so that the student will not spend a disproportionate amount of time on the question. The number of credits assigned to an individual question is clearly indicated. The format of the test is readable and properly spaced. The test has been carefully edited, typed, and collated.

There is effective scoring of a short-answer test when correct and incorrect answers can be quickly checked and counted. If the teacher places a space for answers in the right margin, he will facilitate both the student's answering the

items and his scoring them. Also, in some cases, he may find that a separate answer sheet is more practical than the test paper itself. The use of separate answer sheets, for short-answer tests, is remarkably helpful to the person scoring the tests. It can also result in marked economy, if the tests are regarded as suitable for repeated use, as in different sections or comparable classes.

V. SELECTING THE PROPER TEST QUESTIONS: GENERAL CONSIDERATIONS

Let us return to the teacher who, we assume, has completed the blueprint for the test (Section III) and is now ready to build questions which will measure the specific objectives of the blueprint.

Let us consider the types of questions from which he might choose.

A. Types of Questions

Generally speaking, test experts recognize two general types of test questions: short-answer and essay tests. Each type has its own special characteristics, advantages, and disadvantages. In the following paragraphs, we shall explore each type in terms of these aspects.

1. Short-Answer Items

A short-answer item usually requires the student to complete a statement, to indicate the truth or falsity of a statement, to match certain categories, to select from a group of possible answers the one that completes a statement or answers a question, to rearrange items in a list, or to perform similar operations. In each of these types, the student's answer is in a very brief form: a word, a number, or a symbol. (The term "objective" has been used to describe such items, but frequently in reality it means that the item is easily scored. Thus it is very possible—and often the case—that such an item is not "objective" in the sense of admitting only one answer.) In this publication, the following types of items are considered short-answer items: completion, true-false, matching, and multiple-choice.

On the face of it, a short-answer item appears to have certain advantages. First, a large number of such items can sample many of the objectives of certain types of the unit quickly and efficiently. Second, certain types of short-answer items can be scored or graded more reliably than essay questions. Last, although the fact is not generally recognized by many teachers, certain types of short-answer items *can* measure the student's ability to organize and express ideas (albeit not his own), make judgments, draw inferences, make applications of principles, and react to techniques and subtleties of literature. *Collections* of short-answer items can measure progress toward such objectives.

Along with the advantages, a short-answer item brings with it special problems in construction. These problems often affect the validity, reliability, and other factors of a test. To the inexperienced teacher, a short-answer item appears easy to construct and does not seem to have any of the pitfalls of the essay question. Unfortunately, this is not true. The fact that an item can be easily scored is no guarantee of its validity or its reliability. Constructing a good item

is a challenging task. The teacher will need to use all his skills in expressing ideas clearly and accurately. In a good item, only one correct answer and only one interpretation are possible. His directions must be clear, in terms both of the idea or point he is testing and the manner in which the student is to indicate his answer.

Specific suggestions for building short-answer items are included in Section VI.

Short-answer items should ordinarily not be used to measure the student's creative abilities, or to measure the student's own response—in his own words—except in certain instances, such as brief completion items. Each of these abilities is better measured by an essay question, if it can be measured in a paper and pencil test. The essay question has the needed flexibility of format; the student will need to write at length, rather than merely indicate a correct answer.

2. Essay Questions

The second type of item, namely, the essay question, also has certain advantages. It can be used to measure the student's ability to write (and here we mean to organize and present ideas *of his own*), to respond in his own words to selections in literature, and to do similar tasks in generalizing in his own words. His ability to present ideas in his own words remains one of the chief objectives in most programs in English.

The essay question also has certain disadvantages; it may have low validity, have low reliability, or contain other problems in grading. If the proper wording of the question is essential in a short-answer item, it is even more so in an essay question. There is no place for clumsy or awkward wording; for inaccurate use of words or for "window-dressing," that is, unnecessary information or words. Also, although the item may permit the student to express his ideas, the measurement is not accomplished until the teacher or reader has evaluated what is expressed. Therein lies the weakness—the appraisals tend to be inconsistent.

Suggestions for building effective essay questions are included in Section VII.

B. *Selecting the Proper Type of Item*

At this point, the teacher may well say to himself, "What kind of item is best for my test?" The answer to this question is in two parts: the particular objective he wishes to test and the construction of effective test questions.

The first of these two factors is most frequently overlooked by the teacher. As we have indicated in the sections on the blueprint, the teacher should review the objectives of the blueprint and then bring together the objective and the proper test item. In deciding about the latter, he will ask himself questions such as these: What must the student do to indicate to me that he has attained the objective (that is, the testing objective)? Must he write? Must he only recognize word meanings or be able to use the words? Must he be able to respond to literature read in the unit or to unfamiliar literature? What must the student do to show that he has acquired a certain knowledge? What must he do to show that he has improved in responding to literature, insofar as the specific objectives of the unit are concerned? Must he present his own ideas or merely recognize a

correct idea or response? In short, what processes must he go through that will suggest his having attained the testing objective?¹

Insofar as the second step—building effective items—is concerned, the subsequent sections of this leaflet present suggestions. The teacher should keep in mind that the choice of a particular item is often a form of compromise, in which he accepts certain limitations of an item in order to realize its greater advantages. For example, to test the ability to write, he will probably wish to use a writing exercise of some type, even though it may be somewhat unreliable in its grading. This factor does not mean, of course, that he will not make every attempt to improve the question and its grading by careful wording and limitation as described in Section VII. Similarly, he must work to overcome the problems inherent in other types of test items as well.

VI. BUILDING EFFECTIVE SHORT-ANSWER ITEMS

As indicated in the previous section, in this section we shall examine some of the common types of short-answer items and the special characteristics of each. We shall examine each type in the light of such factors as validity, reliability, ease of scoring, and suggestions for building it.

A. General Considerations

As indicated in Section IV, there are certain objectives for short-answer items, like all test questions, which the teacher must attempt to achieve. An item should have validity and reliability and should be relatively easy to score or grade. Some of the factors which affect these characteristics we shall describe below.

1. The Teacher's Knowledge of the Content and Similar Aspects of the Unit

The teacher will need to recall essential matters of content and to decide which can best be tested by short-answer items. He must also decide what kind of understanding, appreciation, and discrimination can best be tested by such items.

The teacher will keep in mind any errors, insufficiencies in learning, or misconceptions on the part of the student. Through them the teacher can discover bases for many effective items.

2. Clarity and Accuracy of Expression

The teacher will need to express himself accurately and clearly in the short-answer item particularly. He must use the word or words which will best convey his intent and which will have the most meaning to the most students in the class. He will make every word count. He will write clear, accurate statements, questions, and directions. Also, if he is wise, he will avoid the language

¹An interesting analysis of such tasks is included in Bloom's *A Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*. In this publication the editor analyzes test items under various categories such as knowledge, comprehension, analysis, application of principles, and synthesis. Each of these categories is fully explained, together with illustrative test items. See Bibliography, p. 32.

of the textbook; it is too easily recalled by rote learners. Instead, he will rephrase the idea, using different—that is, synonymous or more general or more specific—words which will still convey the desired meaning.

In building a test, the teacher must use standard English. Errors in usage are inexcusable.

3. Imagination

The teacher will need to use his imagination in selecting or devising the kind of test item which will best serve his purpose. He will try to put himself in the place of the student answering the item. He will try to recall the mistakes the student has made in the past regarding the point to be tested. What do the words in the item mean to the student? What process of thought does he go through as he answers the item? Where may he fall down? In a short-answer item, what kind of wrong answer will appeal to the *poor* student so that he will select it? What kind of question will challenge the average student? the superior student?

The teacher must be flexible here as well. He must be willing to reject one type of item in favor of another type, if the latter will better achieve his purpose. For example, if he finds in testing application of principles that a matching-type item is inappropriate, he may shift to a multiple-choice item. Or, if he cannot find a third or fourth wrong answer for a multiple-choice item, he can reduce the number of answers or try to build the item as a True-False or completion item.

B. Types

1. Completion

In a completion item, the student is asked to fill in a blank which completes a statement or answers a question. Below is a sample item:

The name of the magazine in which the section entitled "Life in These United States" appears is _____

A good completion item is difficult to build for one reason particularly: it is inclined toward unreliability of scoring. That is, the item may yield several acceptable answers which may not have been anticipated by the teacher.

a. Uses

Such items have two possible minor advantages: They are largely *recall* items, and, as such, require that the *student* furnish entirely the information required; and, second, they can provide, if used in large numbers, a fairly adequate sample of certain of the unit's objectives. However, they have a severe limitation: they function efficiently only when used to test knowledge of facts. Further, they do not seem to be well adapted to testing the ability to draw inferences, to judge, to reason, or to discriminate.

Below appear typical problems in writing such items.

b. Cautions

(1) Make sure that there is only one answer for the item.

Poor: Robert Louis Stevenson wrote _____

(How many answers are correct here? "About the sea," *Treasure Island*, "novels"?)

Improved: The name of the novel by Robert Louis Stevenson which we read is _____

Restricting the student to the "name of the novel" sharpens the possible answers to only one acceptable one and thereby increases the reliability of the item.

(2) Make sure that all parts of the item function.

Poor: Charles Dickens, one of the outstanding British authors of the 19th Century, wrote a novel about the French Revolution entitled _____

(Is "one of the outstanding British authors of the 19th Century" necessary?)

Improved: Charles Dickens wrote a novel about the French Revolution entitled _____

(3) Do not give any indirect clues to the answer (except in constructing tests for frames in programmed booklets).

Poor: Stephen Leacock is well-known as an _____

(If "essayist" is the correct answer, the student has been guided to it by the word "an," which of course usually appears before a word which begins with a vowel.)

Improved: Stephen Leacock is most famous for writing which type of literature? _____

(Even this item has more than one answer.)

c. Summary

Use completion items with *extreme caution* and then only to test factual knowledges. Their limited applicability to most objectives in a program in English makes them very unprofitable in most tests. Again, do not use them when no *recall* is involved.

2. True-False

This type of item usually appears as below:

The author of *Great Expectations* is Charles Dickens. T F

In this kind of item the student is to circle the correct answer, or place a symbol (T or F [sometimes O]) in the space in the margin.

A modification of this type is the item for correction. In this type, if the student believes the statement to be false, he circles the F and in the blank in the margin places the word or phrase which makes the statement correct.

True-False items are notoriously suspect in one respect: they lend themselves to too much guessing by the student. Inasmuch as there are only two alternatives (True, False), the student has a 50-50 chance of answering an item correctly, even if he is unable to answer it on the basis of knowledge or on any other basis.

a. Uses

To test misconceptions which the student may have in a given area of English, particularly before the beginning of the study of that area, a True-False item is sometimes useful.

In some cases, the item may be used if a great amount of sampling is to be done in a short period of time. For example, to test the student's knowledge of the correct spelling of the word "separate," the item below may serve:

A. Choose the correct spelling:

1. separate 2. seperate A. _____

A large group of such items might be useful, but to a limited extent.

b. Cautions

(1) In an item, test only one idea.

(2) Do not write an item which is partly true and partly false.

(3) Avoid sweeping statements; they tend usually to be false.

Poor: The word "rash" always means "reckless." (Such words as "always," "never," etc., usually contain an element of absoluteness which the student knows is usually not true.)

(4) In modified True-False items, underscore the point being tested. Otherwise, the student will not know where the "nub" of the item lies.

Poor: The author of "The Lady or the Tiger?" is O. Henry. T F _____

(This item is poor because the student does not know whether he is being tested about "The Lady or the Tiger?" or about O. Henry. If he chooses the former, he will change the answer to read "Stockton." If he chooses the latter, what story should he insert in place of "The Lady or the Tiger?"?)

Improved: The author of "The Lady or the Tiger?" is O. Henry.

(There are still only *two* choices in such items.)

c. Summary

The limited applicability of a True-False item for objectives in English makes it very unprofitable in most English tests. In spite of its apparent simplicity, it is extremely difficult to build. There are so many exceptions to universal statements of any type that usually reliability is poor and validity even worse. Therefore, *in most cases*, the teacher will do well to avoid this item type.

3. Matching

In a matching item, the student is asked to pair elements which are arranged in a list of two columns or similar form. In each case, he must associate the element in one column with the proper one in the second column. Below is a sample of this type of item:

Column I contains the names of persons or terms of modern mass media of communication.

Column II is a numbered list of the names of some of these media.

For each item in Column I, select the medium in Column II with which the person or term is most frequently associated, and place its *number* in the space provided.

Column I	Column II	Answers
(a) "Person to Person"	(1) radio	(a)
(b) FM	(2) television	(b)
(c) "Life in These United States"	(3) motion pictures	(c)
(d) Joseph Pulitzer	(4) magazines	(d)
(e) "Camera 3"	(5) newspapers	(e)
(f) "Emmy" awards		(f)
(g) "morgue"		(g)
(h) "Accent on Living"		(h)
(i) scenario		(i)
(j) "Invitation to Learning"		(j)
(k) John Huston		(k)
(l) Walter Lippmann		(l)
(m) Cecil B. DeMille		(m)
(n) Red Smith		(n)
(o) "Postscripts"		(o)

A matching item is usually very easily scored. However, it must be necessarily limited to a *list* of elements which are homogeneous; that is, the elements must have a common characteristic. It requires considerable skill to construct. Also, it takes up considerable space on the test paper.

a. Uses

If well-constructed, a matching item can economically test many kinds of factual knowledge and knowledge of principles, provided that the elements tested have some homogeneity and can be reduced to a list. In the example above, the characteristic common to the elements in both columns is the mass media of communication.

b. Cautions

- (1) Keep elements as homogeneous as possible. Note the variety in the elements in the item below. Note that the list includes the names of definitely unrelated things.

Poor:

Column I	Column II	Answers
(a) Author of <i>A Tale of Two Cities</i>	(1) Boston	(a)
(b) The captain of the H.M.S. <i>Bounty</i>	(2) abridged	(b)
(c) The setting for <i>Johnny Tremain</i>	(3) <i>World Almanac</i>	(c)
(d) A dictionary with all word meanings in it.	(4) Charles Dickens	(d)
(e) A library tool containing sports records	(5) unabridged	(e)
	(6) William Bligh	(f)

- (2) Place the longer statements or terms (elements) in the left-hand column and the shorter elements with which they are to be matched in the right-hand column. This practice facilitates

the student's reading of the item. (Note the example in the previous item above. Although its elements are not homogeneous, the longer elements do appear in the proper column, the left-hand one.)

- (3) To help to insure reliability, eliminate some guessing by the student by including more elements in one column than in the other. This precaution prevents the student from getting an answer correct by the process of eliminating all but the one remaining element. Sometimes, building the item so that an element will be used more than once is a similar precaution.
- (4) Make directions as specific as possible. Indicate the kind of relationship which exists between the elements in the two columns. (In the first sample above, the relationship is between terms associated with the mass media of communication and the names of the mass media.) Indicating that relationship will help the student to understand better what is being asked. Also, if necessary, indicate that an element may be used more than once.
- (5) Put all parts of the question on the same page.

c. Summary

The matching item does have a place in testing if in a given item the elements are homogeneous and can be reduced to a list. In some cases, however, the item may be better cast in the form of a multiple-choice item. Because the matching item has limited flexibility, it functions well in a rather small number of situations.

4. Multiple-Choice

A multiple-choice item usually consists of a partially complete statement or question, called the *stem*, following by a group of alternatives. One of these alternatives is correct; the others (*distractors*) are not. From the group of the alternatives the student is to select the correct answer, placing its number in a space provided in the margin. Below is an example of a multiple-choice item:

The word "intrepid" most nearly means

- (1) helpful (2) happy (3) trustworthy
(4) careful (5) brave

The item may appear in the form of a statement to be completed, as above, or in the form of a question for which the student selects the correct answer from a group of alternatives. Neither of these two forms is inherently superior to the other. The inexperienced teacher should use the question form until he has had sufficient practice in writing such items. He must always keep in mind that clarity of expression, and ease of reading by the student, are essential considerations.

In recent years the multiple-choice has become the type of short-answer item most favorably received by teachers and test experts alike. It is extremely valuable as a test device because of its wide applicability; it can function in almost any kind of test situation.

On the following pages appear different types of multiple-choice items. They give an indication of the large *variety* of testing situations in which multiple-choice items can function well.

Each of the following sentences contains an underlined expression. Below each sentence are four suggested answers. Decide which answer is correct and place its number in the space provided. (5)

- (1) I found that one of the toys *were* broken.
- 1 Correct as is 3 was broke
2 was broken 4 were broke (1)
- (2) After I had finished High School, I went on to college.
- 1 Correct as is
2 High School having been finished,
3 After finishing High School
4 After I had finished high school, (2)
- (3) The store specializes in infants' and childrens' clothing.
- 1 Correct as is 3 infant's and children's
2 infants' and children's 4 infant's and childrens' (3)
- (4) Every man, who breaks the law, should be punished.
- 1 Correct as is
2 man who breaks the law; should
3 man who breaks the law should
4 man, who breaks the law should (4)
- (5) To do a task promptly is better than worrying about doing it.
- 1 Correct as is
2 worrying for doing it
3 to worry about to have done it
4 to worry about doing it (5)

Each of the following questions is about related areas of English. Write in the space provided the *number* of the expression that best completes the statement. (5).

- (1) Which word is most specific? (1) cosmetic (2) luxury (3) lipstick
(4) makeup (1)
- (2) A type of speech which is found largely in a certain geographical area
is called (1) a dialogue (2) an archaism (3) a colloquialism
(4) a dialect (2)
- (3) Famous lines from a play by Shakespeare are most likely to be found
in a book originally compiled by (1) Bartlett (2) Firkins (3) Fowler
(4) Roget (3)
- (4) One reason that the lead of a news story in a newspaper summarizes
the essential information is to (1) aid the hurried reader (2) reduce
the length of subsequent paragraphs (3) save work for the reporter
(4) encourage the reading of advertisements (4)
- (5) Poems are regularly included in (1) *Life* (2) *Reader's Digest*
(3) *Changing Times* (4) *Atlantic* (5)

Many fans are irrational in a ball park, and they will rumpus sometimes even to the point of a near riot over a decision which seems to have been honestly called; but if a raw decision is called against the visiting team, they may even applaud it. There are fans who can be seated behind a pillar but who will yell bloody murder and scream in protest of a decision which they either did not see or saw badly and at too great a distance really to know what happened. During the 1934 World Series, the Detroit fans, in a whooping display of sportsmanship, littered the field; and Ducky Medwick was removed from a ball game at the suggestion of Judge Landis for fear of injury to the Cardinal star or a riot. John McGraw was once reported to have been escorted by police from the ball park in Cincinnati lest he be mobbed. Everyone who has attended many ball games knows that the code of King Arthur's Court does not control the language, the manners, and occasionally, the conduct of every baseball fan of this nation.

The title below that best expresses the ideas of this passage is:

- 1 Riots in the ball park
- 2 The behavior of baseball fans
- 3 Fans and umpires
- 4 The hardships of professional baseball
- 5 Poor sportsmanship on the baseball diamond()

In the last sentence, the author's tone is one of (1) understatement (2) satisfaction (3) admiration (4) resentment (5) optimism ()

In this passage, which word is used in an ironic sense? (1) rumpus (line 2) (2) applaud (line 5) (3) sportsmanship (line 12) (4) escorted (line 16) (5) mobbed (line 17)()

Which word best describes the followers of the home team? (1) impartial (2) sluggish (3) careless (4) bipartisan (5) partisan ..()

Which quality of the spectators at baseball games does the author emphasize?

- (1) nearsightedness
- (2) fair-mindedness
- (3) untidiness
- (4) unreasonableness
- (5) disloyalty()

VOCABULARY

A. Vivid verbs (Best answer)

1. The right verb for the context
 - a. The heavy, fat man *walked* clumsily into the room.
 - (1) walked (2) strode (3) waddled (4) flitted (5) gambolled

B. Connotations of words and discrimination among synonyms on this basis (Best answer)

1. *fatal* (1) serious (2) deadly (3) tragic (4) crushing (5) agonizing
2. Which word has a *positive* connotation in most situations?
 - (1) artificial (2) culpable (3) altruistic (4) deceptive (5) mortified

C. Words at various levels of usage (Best answer)

1. Which word for *policeman* usually appears in writing of a formal nature? (1) cop (2) dick (3) flatfoot (4) constable

D. Unusual (or common) word derivations (Best answer)

1. Which language has produced the word roots "psycho" and "philo"?
 - (1) Latin (2) Arabic (3) Hebrew (4) Greek (5) Russian

E. Elimination of unidiomatic expressions (Best answer)

1. Which one of the following sentences contains an expression which is *not* a correct English idiom?

1. I went up at the library.
2. etc.
3.
4.

- F. Words commonly associated with a given activity (Multiple-response)
1. bowling (1) strike (2) spare (3) chukker (4) round (5) inning
- (For this item, the fact that more than one answer is correct is acceptable because the student's knowledge of a number of meanings is being tested.)

In the following group of literature items for our unit on "Understanding Human Nature," note the *range* of the items. Note that the first item tests factual knowledge, and the subsequent items test appreciation, judgment, response to the techniques of the author, etc. Here observe how flexible multiple-choice items can be. They can test material at a low literal level and are well suited to testing higher mental abilities and skills.

LITERATURE

1. In *The Thread That Runs So True*, the author describes his experiences as a (1) young schoolteacher (2) minister's son (3) tailor's apprentice (4) track champion.
2. In *Julius Caesar*, Brutus's inner conflict is shown in the lines (1) "Speak, hands, for me!" (2) "Cowards die a thousand deaths. . ." (3) "Not that I loved Caesar less. . ." (4) "he doth stride the narrow world like a Colossus."
3. In "They Grind Exceeding Small," Hazen Kinch is shown to be (1) careless (2) greedy (3) lazy (4) sickly.
4. In *Dear Brutus*, the experiences of the persons in the play suggest that (1) some men are influenced more by evil than by good (2) happiness can be too dearly bought (3) human nature is ever the same (4) patriotism is an ever present need of man.
5. In *Julius Caesar*, Caesar's attitude toward Cassius was one of (1) grudging admiration (2) mild approval (3) studied indifference (4) deep mistrust.
6. In "The Last Class," Franz's attitude toward the event in the story was one of (1) joy (2) meek acceptance (3) patient protest (4) keen dismay.
7. The plays of Sir James M. Barrie are characterized by (1) blank verse (2) "choruses" to accentuate action (3) unhappy endings (4) humorous stage directions.

The number of alternatives (usually from three to five) in a typical multiple-choice item tends to decrease the amount of guessing by the student. Furthermore, the number of alternatives tends to reduce the chances of a guess being a correct answer. Also, a multiple-choice item can measure certain skills, knowledge, and abilities more easily than can other short-answer types. It can be very reliable, inasmuch as it is scored by means of a key which admits only one correct answer.

A multiple-choice item has some disadvantages. First, it requires somewhat more time to construct than do other types. Second, it takes up more space on the test than do other item types. Third, it requires a large amount of reading by the student.

a. Uses

A multiple-choice item can effectively be used, as indicated in the examples above, to test knowledge of facts and to test more complex outcomes: appreciation, analysis, and understanding at various levels. Its applicability ranges from an item asking the student to discriminate among five ways of expressing an idea, to apply principles of writing, to understand basic themes of literature, to understand the theme of a selection of poetry, to recognize subtleties in an author's style, to draw inferences, and to perform similar operations. The outstanding characteristic of a multiple-choice item is, then, its *adaptability*.

b. Cautions

(1) Use clear, simple language.

Poor: His early career having been atrophied, Sydney Carton took refuge in (1) drinking (2) eating excessively (3) stealing (4) hoarding money (5) fleeing the country.

(This item presents two vocabulary problems: "atrophied" and "took refuge in." Such expressions may be too difficult for the student.)

Improved: A habit of Sydney Carton's was (1) drinking (2) eating excessively (3) stealing (4) hoarding money (5) fleeing to the country.

(2) Make certain that the stem and the alternatives are easily understood. Be sure that they contain no extraneous information or "general" terms.

Poor: Confronted by the situation that Caesar might become king of Rome, and disturbed by the dreams of his wife, Brutus eventually decided to (1) flee Rome (2) join the conspiracy (3) abandon all hope (4) tell the soothsayer of the plot (5) appeal to Cicero for aid.

(This item is confusing because of the extraneous information in the stem, together with a wordy style.)

Improved: Because of fear of Caesar's power, Brutus decided to (1) flee Rome (2) join the conspiracy (3) abandon all hope (4) tell the soothsayer of the plot (5) appeal to Cicero for aid.

(3) For most items, have one correct answer, and only one correct answer for the item. (Exception: multiple-response items, such as those on page 21 [F].) Although this statement may be a truism, the teacher frequently overlooks its importance. The fault often exists when the teacher tries to build a distractor which comes too close to the correct answer. Below is an example.

Poor: Robert Louis Stevenson was very well known as an author of (1) plays (2) novels (3) chronicles of King Arthur (4) literary criticism (5) poetry.

(Although the teacher intended that number 2 be the correct answer, number 5 is also correct.)

- (4) Have a central problem in the item and place it in the stem beginning of the item. If the teacher selects a specific problem or point and places it in the stem of the item, he concentrates the student's attention upon the point about which he wishes to know. Below are examples of this idea:

Poor: The Federal Communications Commission

- (1) makes rules about radio and television broadcasting
- (2) sets up regulations for the publication of books
- (3) is a group of newspaper publishers
- (4) represents the communications industries in lawsuits
- (5) has jurisdiction over such agencies as the post office.

(In this item, the student is confused as to exactly what the teacher wants to know about the Federal Communications Commission. In the example, the teacher does not direct the student to a specific characteristic or duty of the organization.)

Improved: A function of the Federal Communications Commission is to (1) make rules for radio and television broadcasting (2) set up regulations for the publication of books (3) represent the communications industries in lawsuits (4) have jurisdiction over such agencies as the post office.

- (5) Include as much as possible of the problem in the stem.

Poor: In parliamentary practice, if a member wishes to change a few words of a motion, he may (1) make a motion to limit debate (2) make a motion to adjourn (3) make a motion to appeal from the decision of the chair (4) make a motion to amend.

(Note that "make a motion" appears in each of the alternatives. This fact forces the student to read unnecessarily.)

Improved: In parliamentary practice, if a member wishes to change a few words of a motion, he may make a motion to (1) limit debate (2) adjourn (3) appeal from the decision of the chair (4) amend.

- (6) Make the distractors and the correct answer parallel in form, length, and thought. Also, make them as plausible as possible. Making the distractors parallel with the correct answer, both grammatically and logically, considerably increases the reliability and discrimination of an item. It forces the student to select the answer on the basis of his knowledge and appreciation, not on the basis of other factors.

Poor: The *World Almanac* should be used in (1) locating miscellaneous information (2) searching for pictures (3) preparing a bibliography (4) situations which require good writing and (5) finding the history of the United States.

(In number 4, the answer is not parallel with the others in thought and therefore will probably not distract many of the students.)

Below appears another example:

Poor: Which of the following is an author of novels about the sea?
(1) Melville (2) Bryant (3) Lewis (4) Poe (5) Hawthorne.

(Bryant, as most students will know, did not write novels.)

Improved: An author who wrote novels about the sea was (1) Melville (2) Cather (3) Lewis (4) Poe (5) Hawthorne.

Poor: The form of punctuation which indicates strong feeling is (1) . (2) ? (3) " " (4) : (5) !

(Numbers 3 and 4 are both poor distractors inasmuch as they are not in the same classification as 1, 2, and 5. Therefore, they would be ruled out by the knowing student.)

This item can be improved by reducing the number of alternatives from five to three. There are many times when such a step must be taken.

(7) Place the alternatives *at the end* of the item.

Poor: (1) Stowe (2) Cabell (3) Cooper (4) Poe (5) Norris wrote *Uncle Tom's Cabin*.

(In the above item, a student must read through the five alternatives before he knows what the item is about or exactly what the teacher wishes to know.)

Improved: Who wrote *Uncle Tom's Cabin*?

(1) Stowe (2) Cabell (3) Cooper (4) Poe (5) Norris

(8) Avoid negative or *except* items.

Poor: Which of the following was not written by Charles Dickens?

(1) *Great Expectations* (2) *A Tale of Two Cities* (3) *Little Dorrit* (4) *Ethan Frome* (5) *David Copperfield*

Poor: All of the following wrote novels *except*

(1) Steinbeck (2) London (3) Lewis (4) Cather (5) Shakespeare.

The construction of negative or *except* items creates a serious problem. In the former at least, the point usually can be tested positively. In the latter, there are relatively few situations in which the idea of *except* is really important.

Such item types have a real disadvantage: they usually are not parallel in form to the other items on the test. Consequently, the student is required to make a change in his "mind-set," a process which is confusing to the student.

If such items do appear on a test in a given area, all negative items should be grouped together and all *except* items grouped together.

c. Making an Item Analysis

An interesting method of checking to some extent the validity and reliability of a multiple-choice item is making an item analysis after the test is administered. The analysis can give the teacher a picture of the number of students who chose the correct answer to an item and of the effectiveness of each distractor. An example appears below:

In *Julius Caesar*, before his death Caesar decided to go to the Senate House because of (1) fear of Brutus (2) friendship for Antony (3) fierce pride (4) innate patriotism (5) liking for Cicero.

Below appears a statistical analysis of the item, that is, the manner in which a class of 35 students selected the various alternatives:

	1	2	3	4	5	
Superior	1	0	7	1	1	10
All others	7	4	2	6	6	25
						<u>35</u>

Note that No. 3, the correct answer, was chosen by 70 percent of the superior students. Note that "all others" (other than superior), only 8 percent of whom chose the correct answer, chose the distractors with almost equal randomness. This item is a good item because it discriminates well between the superior and "all other" students and because each distractor functions well.

The teacher can make such a brief tabulation as this for any multiple-choice item and certain other short-answer items. (Often students can assist him in tabulating answers.) If he finds that the distractors function well and that the superior students agree on the correct answer, usually he will have a good item. If, however, he finds that the superior students do not agree on the correct answer or that almost as many "all other" students answer the item correctly as do superior students, he will know that the item is probably a poor one that requires substantial revision.

As indicated in the section on validity (page 9), the teacher must decide what his particular purpose is in using the item.

d. Summary

In spite of its weaknesses and the necessary precautions in building it, a multiple-choice item can test a number of objectives of the program. In recent years, it has become one of the most effective items of a short-answer type.

VII. BUILDING EFFECTIVE ESSAY QUESTIONS

Like short-answer items, essay questions can make a special contribution to the testing of objectives in English. Certain skills and abilities of the student are perhaps better measured by an essay question than by any other type of item. When the teacher wants to measure the student's ability to organize *his own* ideas, his ability to write, to analyze, to judge, to discriminate, to give his own reaction to a work of literature, to create his own literary work, to tell his own experience, all in *his own words*, he may use an essay question.

There are several varieties of essay questions. In the sense in which we use the term here, any exercise which requires the student to write more than a very few words, however briefly, is an essay question. The teacher may overlook the brief essay question, sometimes of only one sentence: "Explain briefly the meaning of the term *semantics*"; "Tell what kind of person General Zaroff in 'The Most Dangerous Game' was," etc. The teacher is probably familiar with the more elaborate questions such as those below in the area of literature.

Essay questions are not easy questions to build. The teacher must allow adequate time to plan such questions very carefully if they are to measure outcomes in their unique way. This statement applies to all essay questions but

particularly to those which are intended to measure such skills as making judgments, analyzing, evaluating, and similar skills. The teacher will need to bring to bear three important qualities: the ability to analyze the objectives of the unit in terms of possible opportunities for writing; the ability to write clear, readable English; and a combination of perceptiveness and imagination in building questions (as in the previous section on short-answer items).

Almost traditionally, teachers use the essay question particularly in two areas, literature and composition, each of which is discussed below. True, such questions can be used in such related areas of English as library usage, parliamentary practice, magazines and newspapers, television, and motion pictures. Occasionally, objectives here can be well measured by essay questions. An analysis of the principles for building literature and composition questions may enable the reader to become familiar with such principles, after which he can apply them to the related areas as well.

Both literature and composition present special problems in the building of good questions. Here are included suggestions for building questions, together with illustrative items.

A. *Essay Questions in Literature*

1. Use the essay question to measure objectives which cannot easily be measured in short-answer form.

Such matters as dates of authors' lives, meanings of literary terms, identification of the settings of stories, and names of important characters in stories are better tested by matching and multiple-choice items and sometimes by completion items. Several types of understanding and appreciation lend themselves well to testing by essay questions. Among these are personal reaction to a work of literature, comparison of themes of different selections, generalizations about human behavior as revealed in literature, and similar problems. In each of these, the student is required to present information about selections he has read and to give specific references to support generalizations which he makes.

Here is an example of our unit on "Understanding Human Nature." The objective being tested is that of "Awareness of the inner conflict of a character."

Example: In our unit we met several people who had serious problems. Some of these problems lay within the persons themselves. Select a person who had such a problem. Indicate what his problem was and the efforts he made to solve it. Give the title and the author of the book in which the person appears.

2. Limit the scope of the question.

A problem often not anticipated by the teacher is that the scope of the question may be much too comprehensive for adequate response by the student. He may need to write entirely too many words in proportion to the total test. Or the question may have vague or ill-defined limits. Here is an example of such a question.

Example: Some of the selections we have read have given us an idea of America. Select two books and in each case show what idea it presented about America.

(This question is subject to several interpretations because it is so general. It also probably lacks reliability; that is, individual students will interpret it in many ways. What phases of American life or ideals should the student discuss? To what extent?)

Here is an improved version of the question.

Improved: In some of the selections we have read during this past year, authors have informed us about America: its ideals, its customs or traditions, its heroes, or its contributions to mankind. Basing your discussion upon one full-length biography and one book of nonfiction, which you identify by author and title, show that you have been informed about one or more of the areas mentioned.

The "Improved" version is much more specific than the "Example." A vague generalization is narrowed to aspects with which the student can deal in the time allotted.

A qualification of this matter is to break down the parts of the directions by placing the letters (a), (b), (c) before each section and to ask the student to answer the parts in the order given.

3. Use clear and accurate directions. Also, eliminate extraneous information.

Below appears an example of a question which has poor directions.

Example: From the novels and plays you have read, choose two books and show that each book has interested you. Books usually do this. Discuss each book thoroughly.

(In its present form, this question is vague and confusing. For example, how many books is the student to write about—two or four? What are the special qualities of interest which he is to discuss? What does "Discuss each book thoroughly" mean? In how many words; also, in what respects? The expression "Books usually do this" is superfluous to the interest of the question, also. Such expressions are "window-dressing" and should be avoided.)

4. In wording a question, use the conceptual level and the level of vocabulary which fit the abilities of the students in the class.

A typical example of the problem of conceptual level is a question which appeared thus on a test for an average English 9 class:

Example: Man's inhumanity to man has been the source of many great literary works. Discuss two such works.

(In this question, it is probable that many of the students in the class were unable to grasp the mature generalization suggested in the first sentence. Apparently the teacher failed to recognize the ability and maturity of the students.)

An example of the problem of vocabulary level appeared on an English 10 test:

Example: Show how the author succinctly gives his views on his subject.

(Unless the student has been taught the meaning of "succinctly," he will probably have difficulty in answering this question. Thus validity is adversely affected.)

5. Define certain terms accurately for the student.

Such terms as "describe," "cite," "discuss," "in detail" must be made clear to the student, either in the question itself or in discussion with the class prior to the test. If the student is unable to grasp the meaning implied in the word, he is at the mercy of the teacher's interpretation. To improve reliability, each term should be defined so as to have the same meaning to all students.

6. In the question, "set the stage" for the specific problem the student is to answer.

In the example below, the first sentence "sets the stage." It gives the student a frame of reference and directs his attention to the general context of the question. The second sentence indicates the specific directions he is to follow in answering the question.

Example: Reading books may enable us to understand better such feelings as fear, love, hatred, remorse, patriotism, or the thrill of danger or conquest. Choose two novels that you have read, and in each show by definite references how the experiences of a person in the book increased your understanding of one or more of the feelings listed. Give titles and authors.

7. Acquaint the student with the factors which you will consider in grading his answer.

There are many factors which the teacher may take into consideration in grading an answer. Specifically, which of these will he consider: requirements of the question? breadth or difficulty of selections chosen? familiarity with selections? ability to generalize? relevance of specific references? technique of composition? For the last, for example, will he consider it, and how? Will all errors count against the student? Only basic errors? Only errors made on work which has been taught? Are some errors more serious than others?

The teacher should answer each of these questions *prior* to the test and inform the student of his decision. In the test itself, then, if necessary, he can refer briefly to the criteria as a whole.

8. Use a series of *short-essay* questions rather than only one question.

This suggestion assumes that the validity and reliability of the test as a whole will be improved, since two questions will provide a better sample of course content than only one.

9. Offer *no choice* of essay questions.

If a student has a choice of one question from among two or more questions, in effect he is not taking the same test as the student who selects another one of the choices offered. Despite attempts to make the various questions equivalent to one another, the teacher is, in effect, discriminating against some students. If a student is not forced to answer a question in an area in which he has little knowledge or understanding, but instead can elect another question, in effect he is not taking the same test as his fellow students.

10. Improve the procedure for grading answers.

a. Set up a key for use in the grading of answers.

(1) Write the best possible answer you can for the question.

- (2) Identify the qualities to be expected of a superior answer; of an average answer; of a poor answer.
- (3) Read several student answers to the question and revise the "qualities" list in 2 above.
- (4) Assign weights to each part of the key, in the manner indicated below for an objective of our sample unit:

Selection meets the requirement of the question	1
Answer reveals student's grasp of essential of content listed by question	3
Answer reveals ability to generalize about the character	4
Answer reveals adequate technique of composition	2
	10

- b. Rate all answers to a question before proceeding to grade answers to the next question.

Following this procedure tends to reduce the "halo effect" caused by the student's performance on questions previous to the one being rated—and on previous tests.

- c. In grading, consider only those factors which give evidence of the extent to which the student has attained the objective set up for the work in literature.

B. Essay Questions in Composition

The essay question in composition often produces uncertainty on the part of both the teacher and the student. Some of these areas of uncertainty we shall now consider.

The first of these is the definition of the term *composition*. Secure as he may be in his concept of the term, the teacher may find that in actual practice he fluctuates. For example, in his teaching he may have a general view of the area of *mechanics*. To be fair to the student, he must define the term exactly if it affects his grading of a composition. Similarly he should define such elements as *content*, *organization*, etc. Will he require that the student have an *introduction*, *body*, and *conclusion*? If he cannot define these generalized terms, how can he construct a valid test question to measure achievement in them?

For example, in some situations the objective "to write a clearly developed paragraph" will maintain validity to the desired degree. In other situations, with a different purpose on the part of the teacher, the objective "to write a paragraph in rebuttal of the chief argument of the speaker in last week's assembly program" will express his objective more accurately. Thus, the teacher will take this first, essential step in building the question, defining exactly what he wishes to measure.

His next step will be to devise an exercise which will reveal the student's achievement of the desired outcome. Each exercise, as in the essay question in literature, will be limited in scope. It must be one that all students can write on, yet narrow enough for answers to be "controlled." A broad generalization such as "Show the effects of a sense of duty in the world today" will probably result

in many interpretations. On the other hand, such a topic as "Explain *one* contribution of America to world culture in the area of entertainment" will probably limit the scope of the answers. So will establishing a word limit for the answer. Again, as in essay questions in literature, the teacher will wish to focus the topic sharply and to make its point clear to himself—and to the students!

For the grading of composition, as yet no completely effective system has been devised. Among test experts, the pendulum appears to swing back and forth from a restricted list of qualities of an answer, together with a certain number of points for each quality, to a practice of "rate the answer as a whole."

Some practices also involve ranking of students' answers. In the latter the teacher places the students' answers into three or more piles representing varying achievement (Superior, Good, Average, Weak, Poor). Such a practice, it is claimed, improves the reliability of the grading when more precise grading is not possible or significant.

Regardless of which scheme is used, however, the teacher may find a key for grading answers a valuable asset. In the case of some tests, the key is a series of generally stated competencies which the student must demonstrate; in other tests, the key is a list of the competencies which will be broken down into specific credits as in our sample blueprint. In an effective key, the qualities of a good answer are defined, and the student is graded *at least* on the basis of these qualities.

To summarize, the teacher can improve his essay questions in composition by

1. Defining specifically what he wishes to measure.
2. Setting up a specific writing situation, the successful completion of which will reveal the student's achievement.
3. Informing the student of what is expected in an answer.
4. Using as reliable as possible measures of grading essays.
 - a. Providing a key.
 - b. Reading all answers to a question before proceeding to the next question.

Further suggestions for grading essays are included in several sources in the Bibliography, notably the books by Hawkes *et al.*, Lindvall, and Weitzman.

VIII. REVIEWING THE TEST

Before he presents his test to the student, the teacher will wish to give it a final review, both before it is typed and collated and afterwards.

Here is a list of criteria for him to consider:

A. *The Test as a Whole*

1. Are all important objectives of the unit tested?
2. Do the questions represent a good sample of the unit's content and emphases? Are weights and credits appropriate?
3. If the unit's objectives include the ability to judge, discriminate, respond to author's tone, and similar objectives, is provision made in the test for testing them?

4. Are questions clearly stated, in words which convey the exact meaning intended?
5. Are provisions made so that answers can be scored or graded as reliably as possible?

B. *Types of Questions*

1. For each testing objective, is the most appropriate item used to test it?
2. For short-answer items:
 - a. Is there a correct answer, and, where applicable, only one correct answer?
 - b. Are questions clearly stated?
 - c. Are directions for answering clearly stated?
 - d. Is "window-dressing" avoided?
 - e. Is *one central* problem tested by the item?
 - f. In multiple-choice items, do all distractors function? Is each one parallel to the correct answer in thought and form? logical? plausible?
 - g. Are indirect clues to the correct answer avoided?
3. For essay questions:
 - a. Is the testing objective best measured by an essay question?
 - b. Is the scope of the question properly limited?
 - c. Are the vocabulary and conceptual level used appropriate for the group being tested?
 - d. Are introductory statements and directions completely clear?
 - e. Are specialized terms such as "cite," "explain," etc., defined so that the student is clear how their meaning affects his answer?
 - f. In general, is more than one such question included, to improve validity?
 - g. Is a carefully worked out and tried out key provided for grading answers?

C. *The Test Paper*

1. Identification: Is there space for the student to place his name, grade, and section?
2. Directions: Are they clear and complete?
3. Numbering and credits: Are all questions numbered? Are amounts of credits in a prominent position? Are subdivided credits clearly indicated?
4. Space for answers: Is it sufficient, particularly for completion and essay answers? Is additional paper provided? Is a separate answer sheet feasible?
5. Scoring: For short-answer items, is there a space in the right margin, where possible, for the student to put his answer?
6. Grading: Are a key and sample answers available and used properly?
7. Editing: Is the copy free of typing and mechanical errors, and attractive?
8. Time: Is the test as a whole too long? too short?

D. Evaluating the Test

Once the test has been administered, its effectiveness can be determined by using certain procedures.

One of these procedures is making an item analysis, as shown on pages 24-25. A second is estimating the difficulty of the test. In this connection, an item answered correctly by nearly 100 percent of the class is useless; so is one answered correctly by only 5 percent of the class. A third useful technique is to have the students comment on items and questions, especially their wording, and the effect of the distractors in short-answer items. The teacher may also wish to check the students' performance on the test with standardized tests or subtests which measure the same objectives as those of his test.

The teacher will also benefit from the experience of other members of his English department, of his statement department of education, of commercial testing organizations, and of other organizations such as colleges and universities interested in effective tests. When the teacher recalls that nationwide testing services, which conduct tests in schools, are continually revising and refining test items and questions, he will see the great value of constant reevaluation and co-operation with others in building better tests.

BIBLIOGRAPHY

(Sources listed below are available in college or university libraries.)

1. Bloom, Benjamin S., et al. *A Taxonomy of Educational Objectives. Handbook I, Cognitive Domain*. New York: David McKay, 1956. \$1.95.
2. Gerberich, J. R. *Specimen Objective Test Items: A Guide to Achievement Test Construction*. New York: David McKay, 1956. \$4.75.
3. Hawkes, Herbert E.; Lindquist, E. F.; and Mann, C. R. *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin, 1936. O.P.
4. Lindvall, C. M. *Testing and Evaluation: An Introduction*. New York: Harcourt, Brace & World, 1961. \$3.60.
5. National Society for the Study of Education. *The Measurement of Understanding*. 45th Yearbook, Part I. Chicago: University of Chicago Press, 1946. \$2.25.
6. Thorndike, Robert L., and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education*. New York: John Wiley and Sons, 1961. \$7.75.
7. Tinkelman, Sherman N. *Improving the Classroom Test*. Albany: The University of the State of New York Press, 1956. \$.30.
8. Weitzman, Ellis, and McNamara, Walter J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. O.P.